From Ignorance to Evidence? The Use of Programme Evaluation in Conservation: Evidence from a Delphi Survey of Conservation Experts.

Hannah Fay Curzon[1] and Andreas Kontoleon[2]*

Abstract

Persistent gaps in the evidence base regarding the performance of conservation policies has put pressure on the conservation policy field to adopt 'best practice' programme evaluation methods. These are methods that account for the counterfactual and are able to attribute causality between a conservation policy and specific observable environmental and social impacts. Despite this pressure, use of such methods continues to be rare. This paper uses the Delphi technique to provide the first systematic assessment of the reasons behind the apparent hesitation of conservation practitioners to adopt rigorous policy impact evaluation methods. The Delphi study consisted of two online questionnaires conducted on conservation policy experts. The results presented confirm that the use of rigorous impact evaluation methods in conservation is still very limited but this, crucially, is not because conservationists are ignorant of these methods or their advantages. In fact, considerable effort is being made to develop and improve evidence standards but these efforts have largely been thwarted by large financial and time related constraints that mean even elementary evaluations are hard to achieve. The results from this Delphi study allow us to provide more realistic recommendations on how impact evaluation studies can be more widely embraced and implemented in conservation practice.

[1,2]Department of Land Economy, University of Cambridge, 19 Silver Street, Cambridge, CB3 9EP
*Corresponding Author: Prof Andreas Kontoleon (E-mail: ak219@cam.ac.uk; Tel: +44 1223 339773)

1. Introduction

Conservation practitioners and policy-makers need credible information regarding the performance of conservation interventions in order to ensure that scarce funds are not wasted on ineffective policies (Sutherland et al., 2004; Stem et al., 2005; Botrill, Hockings & Possingham, 2011). There have been numerous calls for the conservation policy field to adopt 'best practice' or 'rigorous' programme evaluation methods (e.g. Ferraro & Pattanayak, 2006; Ferraro, 2009). These methods focus on the use of experimental and quasi-experimental evaluation designs that can be used to credibly measure 'counterfactual' outcomes. It is argued that establishing this counterfactual is critical to being able to unambiguously isolate the impacts of policy interventions so as to get an unbiased estimate of a programme's performance (Berry et al., 2012).

Despite these calls, there are still large gaps in the hard fact evidence base regarding the performance of conservation policies. Several reviews have documented the paucity of formal evaluations studies on conservation policies using experimental and quasi-

experimental methods (e.g., Pattanayak, Wunder & Ferraro, 2010; Blackman, 2012; Miteva et al., 2012; Adhikari & Agrawal, 2013; Roe, Greig-gran & Mohammed, 2013; Zheng et al., 2013; Alcorn, 2014; Cowling, 2014; Samii et al., 2014). This body of work has found that though monitoring and evaluation data (which only documents trends and changes in variables) is abundant and routinely collected, formal evaluation studies (which identify the causal links between a policy and specific conservation outcomes) are highly scarce

Although the inherent financial, temporal, logistical, and sometimes ethical, challenges of conducting rigorous evaluations have been discussed in the literature, it is still conjectured that one of the main reasons for the limited use of policy evaluation methods is not through a lack of opportunity and resources but, instead, due to a lack of awareness, understanding and appreciation of the need for counterfactual thinking within the conservation policy field (e.g., Ferraro 2006, 2009, 2012). Such assertions are, however, largely unsupported by any kind of formative assessment of the rationale behind conservation evaluation decisions in practice and thus risk being inaccurate and out-of-date. Arguably, in order to obtain a more comprehensive understanding of the underlying reasons for the documented gaps in the evidence base, it is necessary to draw on the knowledge and experience of the actual decision-makers and practitioners working in the conservation policy field. The present study aims to fill this research gap by being the first to systematically ascertain information from experts working in conservation as to their stance with respect to the usefulness, practicality, desirability and prospects of using formal policy evaluation methods. For this purpose, our study uses the Delphi technique, an iterative survey-based research method, which allows for a systematic assessment of the conservation sector's actual knowledge, appreciation, and experience with such methods. As a result, our study will be able to more critically evaluate the commonly made assertions found in several past reviews that the conservation sector is averse to impact evaluation. Lastly, the study will provide policy relevant information on how to more rigorously determine the needs, opportunities and barriers to using 'best practice' methods to evaluate the impact of conservation interventions. These findings could significantly contribute to improving our understanding of the conservation sector's approach to evaluation and how far conservation organisations represented in this study are thinking counterfactually, thus providing a more accurate and informed assessment of the real reasons for the gaps in the evidence base.

This paper proceeds as follows: Section 2 provides some of the common critical assertions found in review literature on the paucity of impact evaluation work in the conservation field. This is followed by the rationale for this study and the specific research questions we address. Section 3 outlines the research methodology as applied in the Delphi study. The results of the study are then presented in Section 4 and are discussed and summarised in Section 5. The survey instruments that were used appear as Supplementary Materials (Appendices SM1-SM4). More details specifically on the methods used can also be found in a SM1 (Technical Annex).

2. Impact evaluation in conservation policy

2.1. The impact evaluation revolution in science

Programme evaluation is fundamentally a process of making inferences about an unobserved counterfactual outcome, i.e., what would have happened in the absence of the intervention, programme or policy. (Ferraro & Pattanayak, 2006). Without this 'counterfactual analysis' it is impossible to know how far impacts are the result of the intervention and not due to other confounding factors or biases (White, 2006; Khandker, Koolwal & Samad, 2010). However, as the counterfactual cannot be observed, the main challenge of impact evaluation is to find or construct an appropriate counterfactual in the light of the missing data.

Two common approaches to evaluation that have been used in the conservation policy field are before-after and with-without comparisons, i.e., comparisons of outcomes before and after an intervention and comparisons of outcomes in areas with and without exposure to the intervention. As before-after comparisons do not control for other time varying factors, and with-without comparisons do not control for selection bias, both methods lead to biased estimates of impacts (Khandker, Koolwal & Samad, 2010). More rigorous approaches that can be used to solve the problem of selection bias and establish a credible counterfactual broadly fall into two categories (Khandker, Koowal & Samad, 2010). The first relies on data obtained from randomised controlled evaluations or trials (i.e. RCTs) which randomly assign study subjects into treatment and control groups. The data is collected before and after the policy leading to the so-called Before-After-Control-Impact (or BACI) design which is widely regarded as the 'gold-standard' in programme evaluation (Frondel & Schmidt, 2005; Duflo, Glennerster & Kremer, 2008; Greenstone & Gayer, 2009). By randomly allocating treatment and control groups across eligible sample units, units that do not receive the treatment will be a valid comparison group for those that did since there should be no systematic differences between their characteristics (Rossi & Freeman, 1993).

When randomisation of the treatment is not possible, the second-best option is to rely on observational data of two samples of subjects, one that has been exposed to a policy (or treatment) and others that have not. Then practitioners use quasi-experimental statistical methods (such as propensity score matching and difference-in difference estimation) to create comparison groups that are valid under a set of underlying assumptions about the nature of potential selection bias in programme targeting and participation (Khandker, Koowal & Samad, 2010). While these econometric methods are well-developed and firmly grounded in theory and statistics, the identifying assumptions are not always directly testable, and the validity of any particular study depends instead on how convincing the assumptions appear (Duflo, Glennerster & Kremer, 2008).

The call for the use of formal impact evaluation methods that address the issue of the counterfactual is part of a broader movement towards evidence-based policy making (Gertler et al., 2011) that was first experienced in medicine in the second half of the twentieth century (Pullin et al., 2004). The resulting paradigm shift from ‘experience-based’ to 'evidence-based’ practice that emphasized the use of clinical experiments and systematic reviews (Pullin & Knight, 2001; Stevens et al., 2001) completely

revolutionised medical practice. This 'effectiveness revolution' became the archetypal method for evaluation and primary research and spread to other social policy fields such as public health, education and international development who started to build randomised evaluations into their programmes recognising the need for convincing and comprehensive evidence that could be used to inform policy making and improve the allocation on government resources ((Pullin & Knight, 2004; Pullin et al., 2004; Gertler et al., 2011).

2.2. Impact evaluation in conservation policy

In contrast, the field of conservation policy did not experience the same 'effectiveness revolution' and even by the beginning of the twenty-first century the evaluation of conservation programmes continued to be rare (Kleiman et al., 2000). One of the main conclusions stemming from a global review of the evidence base known as the 'Millennium Ecosystem Assessment,' was that '[f]ew well-designed empirical analyses assess even the most common biodiversity conservation measures' (MEA, 2005, p.122). Indeed, it was widely acknowledged at the time that conservation was still largely an experience-based practice that depended on intuition and anecdote to guide the design of conservation investments as opposed to empirical evaluations (Kleiman et al., 2000; Pullin & Knight, 2001; Salafsky et al., 2002; Salafsky & Margoluis, 2003; Pullin et al., 2004; Sutherland et al., 2004). While these studies advocated the need for evidence-based conservation, interest in impact evaluation *per se* did not emerge in the conservation policy field until the mid to late 2000s (Frondel & Schmidt, 2005; Ferraro & Pattanak, 2006; Ferraro, 2009; Greenstone & Gayer, 2009; Pattanayak, Wunder & Ferraro, 2010). As a result, the amount of literature on environmental impact evaluation is still limited.

Ferraro and Pattanayak's 2006 paper was one of the first to call for rigorous empirical evaluation of conservation polices. The authors argued that while conservation projects had increasingly focused on 'monitoring and evaluation' since the 1990s, 'rigorous measurement of the counterfactual in the conservation literature was non-existent' (Ferraro & Pattanayak, 2006, p.483) which had not only left conservation policy lagging behind most other policy fields but had also created a large gap in the evidence base regarding the effectiveness of even the most common conservation interventions (Ferraro & Pattanayak, 2006). The authors argued that:

> If *any* progress is to be made in stemming the global decline of biodiversity, the field of conservation policy research must adopt state-of-the-art program evaluation methods to determine what works and when.
>
> (Ferraro & Pattanayak, 2006, p.482)

Particular emphasis was placed on the need for more experimental and quasi-experimental evaluations in the conservation sector on the basis that nearly all environmental programmes have hidden confounders which means non-rigorous evaluation approaches will, in most cases, lead to biased estimates of programme effectiveness. While the authors recognised the methodological challenges to using these approaches, they argued that there were still 'substantial opportunities [in the conservation policy field] to elucidate causal relationships through experimental and quasi-experimental designs' (Ferraro, 2009, p.76). In the same year Greenstone and Gayer's (2009) paper also stressed the need for policy makers to place greater emphasis

on credible empirical approaches. Again, randomised evaluations were recognised as the ideal way to achieve this but the paper also demonstrated the validity of quasi-experiments as an appealing alternative.

Since then, subsequent review papers on the environmental and social effectiveness of conservation policies have all reached a similar conclusion, namely that evaluation studies that construct credible counterfactuals are scarce and that there is a reluctance and hesitancy to undertake such exercises within policy circles (e.g., Blackman & Rivera, 2010; Blackman, 2012; Miteva et al., 2012; Roe, Grieg-gran & Mohammed, 2013; Samii et al., 2014; Fisher et al., 2014; Baylis et al., 2015; McKinnon et al., 2015). In addition to highlighting the need for more rigorous evaluations, the review studies mentioned above have tried to identify and characterise some of the difficulties and potential barriers to implementing experimental and quasi-experimental designs in an attempt to provide some explanation for the limited use of these methods. Some of the barriers mentioned include missing baselines, long time-lags between intervention and impacts, complex spill-over effects, ethical considerations, lack of funding for evaluations and lack of time to update evaluation best-practice guidelines. Whilst these barriers to formal evaluation are recognised as being particularly pervasive in the conservation policy field, the 'real' reasons behind the limited amount of credible studies is yet to be formally assessed and thus remains contested. For example, Ferraro and his colleague's argument is that the limited use of rigorous evaluation in conservation is due to lack of awareness and understanding of state-of-the-art programme evaluation methods, and a lack of appreciation for the biases in standard evaluation techniques (Ferraro & Pattanayak, 2006; Ferraro, 2009). According to Ferraro:

> Environmental scientists and practitioners often assume that evaluation is simply an act of taking a careful look at the monitoring data. If the indicator improves, a program is deemed to be "working." If the indicator worsens, one infers the program is "failing."
>
> (Ferraro, 2009, p.77)

Yet, such assertions are largely unsubstantiated by any kind of formative assessment of the rationale behind conservation evaluation decisions in practice and the apparent hesitancy in using formal impact evaluation methods. Understanding the reasons for the gaps in the evidence base requires examining the merits, need for, and challenges of impact evaluation from the perspective of the policy-makers and the conservation practitioners on the ground. Relying only on reviewing existing and accessible impact evaluation studies (either published in journals or in grey literature) cannot adequately shed light as to how the conservation sector views these methods nor (more importantly) *why* they have not been espoused to the same degree observed in other social policy areas (such as development aid and health care). Indeed, more recent studies suggest that current understanding of programme evaluation in conservation may now be out-of-date, again, supporting the need for a more pragmatic assessment of the evaluation process. For instance, a recent study by McKinnon and her colleagues (2015) argues are that 'CNGOs [conservation non-government organisations] are increasingly engaged with impact evaluation' (p. 3) and that 'investment in producing and commissioning impact evaluations among CNGOs is therefore growing…' (p. 3) but that 'little attention has been given to the organisational arrangements and processes

by which these evaluations occur' (p. 2), again supporting the need for further exploration in this area.

2.3. Aims and research questions

The aim of this paper is to address this gap in the systematic assessment of the use of programme evaluation approaches in conservation by employing an expert-panel based research method which allows us to assess past criticisms of the conservation policy field. Obtaining a more comprehensive view of the situation necessitates seeking perspectives of policy-makers and practitioners working in the conservation sector who have first-hand experience and knowledge of the fields' evaluation aims and techniques. For this reason, this study will use a panel survey of experts working in the conservation policy field to address the study's research questions. Specifically, this study will investigate the following research questions:

RQ1. How important are experimental and quasi-experimental evaluation methods?
RQ2. What are regarded as 'best practice' evaluation methods?
RQ3. To what extent are conservation organisations using experimental and quasi-experimental evaluation methods?
RQ4. What are the most significant reasons for the limited use of experimental and quasi-experimental methods?
RQ5. What efforts are being made to improve evaluation standards?

Addressing these questions will enable us to provide an 'insiders' perspective of how far the sample of conservation organisations represented in this study are actually embracing state-of-the-art evaluation methods. By drawing on the expert knowledge of individuals working on large-scale conservation projects, we expect the insights from this study to be germane to other organisations in the conservation sector.

3. Methods

3.1. The Delphi method

We employ the Delphi method which is a survey-based research method that is able to ascertain the opinions of a purposively selected panel of experts (Hasson, Keeney & McKenna, 2000). Using a series of iterative questionnaires, the Delphi method facilitates structured communication between the experts effectively allowing the group, as a whole, to deal with a complex problem (Linstone & Turnoff, 1975) and to ultimately reach a consensus or convergence in opinion (Angus et al., 2003). The Delphi method has been employed in numerous disciplines including planning, social policy, nursing and information systems research, but also more recently in conservation and natural resource management (e.g., Hess & King, 2002; Oliver, 2002; MacMillan & Marshall, 2006; Geneletti, 2008; Orsi et al., 2011).

3.2. Selection of the expert panel

Unlike traditional surveys, a Delphi survey requires a sample of qualified experts that

have a deep understanding of the issues. Subsequently, rigorous selection of the panel is one of the most critical requirements of any Delphi study (Okoli & Pawlowski, 2004). The experts involved in this Delphi study were primarily identified in three ways: through involvement in projects run by the Cambridge Conservation Initiative (CCI) (a unique collaboration between the University of Cambridge and many of the largest biodiversity conservation organisations in the world (including the World Conservation Monitoring Centre (WCMC-UNEP) and Fauna and Flora International (FFI); through membership of BIOECON (a network of social scientists and policymakers working on conservation policy); and finally, by browsing staff profiles on the websites of national and international conservation NGOs and government agencies such as the World Wildlife Foundation (WWF), The Rainforest Alliance, The Department for Environment, Food and Rural Affairs (DEFRA), The Royal Society for the Protection of Birds (RSPB) and Natural England to name some examples. As there was a need for all of the panellists to speak English, organisations were predominately located/operating in the UK, Europe or North America.

These sources initially produced a list of approximately 1,600 potential candidates. To be selected for the expert panel participants had to work for a conservation organisation as a policy advisor (designing and producing conservation policy interventions) or, have experience working as a conservation practitioner (implementing and evaluating conservation policy interventions on the ground). Conservation researchers working purely in an academic capacity, i.e., not involved with the actual design and implementation of large-scale conservation policies, were excluded from the Delphi panel selection.

Using web searches to obtain biographical information, approximately 300 individuals were identified as meeting the necessary criteria and for which the necessary contact information could be sourced. Individuals were then sorted in order of preference based on their level of experience and expertise. The most preferred candidates were individuals in more senior positions, such as programme managers or officers, and individuals working in monitoring and evaluation (M&E) divisions or those known to have specialist knowledge in conservation effectiveness or evaluation based on their career history. Finally, with an expected 10% response rate we invited the top 200 experts to participate in the Delphi study in order to obtain a sample of a minimum of 20 participants which is in accord with Delphi method best practice guidelines (see SM1. for details).

3.3. Structure of the Delphi process

The Delphi study took place in June 2014 and consisted of two online questionnaires each taking approximately 20 minutes to complete. In our case, an *a priori* decision was made to have two rounds of questions due to the time available and to avoid the risk of sample fatigue. The questionnaires were created using the online survey software 'Qualtrics.' Participants had eight days and nine days (with a week in between) to respond to the first and second questionnaire, respectively. To enhance response rates multiple follow-up email reminders were sent.

To better address the study's research questions, the first round (R1) of questions was structured into two parts (see SM3.). Questions in part one of R1 were designed to address research questions one and two, i.e., the importance of rigorous evaluation methods as well as the panel's perspective on what they considered to be 'best practice' in conservation evaluation. Questions in part two of R1 were designed to address research questions three, four and five, i.e., how far different evaluation methods are actually being used in practice, the panel's perspective on what they considered to be the most significant barriers to using rigorous evaluation methods and how far attempts were being made to improve evaluation and evidence standards in conservation. The first questionnaire was accompanied by a two-page document which introduced respondents to the concept of the 'counterfactual' and outlined different approaches to evaluating conservation interventions (see SM2.). Particular attention was given to the definition of experiments and quasi-experiments in comparison to simple 'before-after' or 'with-without' approaches.

Preparation for the second questionnaire (round two or R2) was devised based on the responses from R1 and was designed to provide a more detailed judgement on the issues therein. Following standard Delphi method best-practice guidelines we largely re-iterated the questions asked in R1 but now also included additional options for the experts to choose from based on the answers provided in R1 (see R2 survey in SM2.). This allows panellists the opportunity to re-evaluate their opinions in light of the new options and information available. In accordance with the Delphi methodology, the panellists were provided with the results from R1 to aid the re-assessment procedure (Angus et al., 2003). See SM1-SM4 for detailed information and justification regarding the structure of Delphi survey, the differences between the two rounds of questions and copies of the survey instruments used in R1 and R2 of the study.

3.4. Composition of the Delphi panel

Table 1 summarises the composition of the Delphi Panel. A total of 45 experts agreed to participate in the study and completed the first round of questions. While the initial response rate was relatively modest (24 %) the number of responses achieved was well above the minimum target of the 20 participants needed for the study. In contrast, there was a much higher response rate to R2 of the study (80 %) as only nine experts dropped out of the study leaving 36 participants (Table 1). For both R1 and R2 of the study there was a relatively even split between the number of males and females serving on the panel (Table 1). On average panellists had worked in conservation for 11 - 20 years, indicating a relatively high level of experience. A total of 24 conservation organisations were represented by the panel members in R1 decreasing only slightly in number to 21 in R2 (Table 1). These organisations included a mixture of national and international NGOs, UK government agencies, major conservation research institutes and international development organisations (Table 1). The majority of the panel were found to be trained conservation scientists, environmental economists or experienced programme officers.

[INSERT TABLE 1]

4. Results

4.1. Delphi survey: round one

Table 2 provides the summary statistics of the responses to seven attitudinal questions posed to panellists in R1 of the study. There was found to be a consensus (75 % or more of the panel agreed) of opinion amongst the panellists for six of these questions which were subsequently omitted from R2 of the study.

[INSERT TABLE 2]

Unsurprisingly, nearly all panel members agreed that evaluations (in general) are essential to building the evidence base. More interestingly, it was also found that a very high percentage of panel members also agreed that the use of experimental and quasi-experimental evaluation methods was particularly important. Yet, over three quarters of the panel had the sense that when it comes to evaluating the success of its interventions, conservation is most likely still lagging behind other policy fields. It is not therefore surprising that the vast majority of the panel members agreed that attempts to measure the outcomes and impacts of conservation interventions using programme evaluation methods were only made occasionally, agreeing also that of these evaluations only some, as opposed to most or all, used experimental or quasi-experimental designs. That said, the vast majority of the panel also believed that there had been at least a slight increase in the use of these methods in conservation over the last few years (Table 2). In contrast, there was found to be less agreement amongst the panel as regards to whether or not conservation organisations are working hard to improve evaluation standards In this case, only 58 % of the panel agreed with this statement. One panel member commented that they would have answered the question differently had the questions asked them to consider just their own organisation and not conservation organisations generally. Subsequently, a re-phrased version of this question was included in R2 of the questionnaire for re-assessment by the panel.

In order to assess whether or not there is any consensus amongst the sample of conservation organisations represented by the panel as regards the best standard of evidence, panel members were also asked to choose which evaluation design, from a list of four options, came closest to what they considered to be 'best practice' in terms of: (a) desirability and; (b) feasibility. Overall, there was found to be no consensus amongst the panellists who provided a wide range of answers for both questions.

As no consensus was reached these 'best practice' questions were repeated in R2. Differently to R1, four additional options were included which were devised based on the 19 suggestions made by the panel as these represented totally different approaches (see Figure 1). This time panellists had the option to choose one or two answers. The questions were also rephrased to emphasise the distinction between what methods were most desirable *in theory* and what methods are actually most commonly undertaken *in practice* within the conservation community. A further modification was that there was no distinction made between methods using 'quantitative' and 'qualitative' data as this distinction, which was made in R1, appeared to have clouded consensus regarding the preferred method.

[INSERT FIGURE 1]

While there was again found to be no overall consensus in R2, panellists appeared to react to the information from other experts as there was less disparity between the group's answers suggesting there was some convergence in opinion. For instance, in terms of desirability, the three most suggested answers provided by the panel were roughly evenly split between an 'experimental design,' (33 %), a 'BACI experimental design' (39 %) and a 'BACI-quasi-experimental design,' (39 %) (Figure 1). One panellists qualified their answer by stating:

> While seeking to codify best practices in conservation impact evaluation is important, we need to recognise that appropriate evaluation design is context dependent, shaped by the level of uncertainty involved in the intervention, the data available, and the needs of decision-makers.

Another panellist added:

> The problem is not about deciding what the bars are and which bar is necessary to evaluate intervention effectiveness. Most of the program leaders get it. The challenges sit with the reality of implementing even some of the lower standard evaluation designs. In many circumstances they are simply not feasible.

In other words, the three favoured designs were all either experimental or quasi-experimental suggesting that these were definitely preferred by the majority of the panel compared to the other non-experimental options. The opposite was found in terms of what methods were said to be used most commonly in practice. This time the three most suggested answers were a 'simple before-after design' (61 %), a 'simple with-without design' (42 %), or 'other,' suggested by 25 % of the panel (Figure 1). Many of the panellists that selected 'other' specified that simple before-after designs were most used in practice but that they most commonly had 'small' samples of treatment groups.

In order to assess whether or not there is any consensus regarding the reasons for the limited use of rigorous impact evaluation methods, panel members were asked to select what they considered to be the five most significant barriers to implementing these methods from a list of 14 options. Panellists also had the option to specify an alternative suggestion. The five barriers most suggested by the panel in R1 were 'lack of funding', 'availability of a baseline', 'time constraint,' 'lack of forward planning' and the 'availability of suitable control group,' (see Figure 2). This question was repeated in R2 of the study to try and build consensus. Building on R1, the panel had a choice of four additional options that were added to the list based on the alternative suggestions provided by experts in R1. Despite the additional options, the five most suggested barriers remained the same (Figure 2). However, this time there was found to be some consensus in opinion amongst the panel as 78 % of the experts concurred that 'lack of funding' and 'time constraint' were two of the most significant barriers. For instance, one panellist stated that:

> Funding is so short-term and funder requirements/interest so inconsistent, that it is basically impossible to develop consistent monitoring programs and to maintain consistent strategies through time.

Another panellist commented that:

> The available funds do not even permit even the most basic level of monitoring.

[INSERT FIGURE 2]

4.2. Delphi survey: round two

R2 of the Delphi study included several new questions. These questions were based on the general comments and feedback provided by the experts in R1. They were designed to provide a deeper insight into evaluation practices within the sample of conservation organisations represented by the expert panel.

In order to better understand reasons for gaps in the evidence base, the panel were presented with a series of plausible explanations (devised from comments in R1) and then asked to score how far they agreed or disagreed with each explanation using a five-point Likert-scale (Table 3). There was found to be a clear consensus in opinion amongst the panellists for two of the four explanations with exactly three quarters of the experts agreeing that 'gaps in the evidence base have less to do with the nature of the field and more to do with a lack of incentives and/or funding/resources,' and that gaps 'can mainly be attributed to a lack of funding and/or resources and not because impact evaluation is not valued in the conservation policy field'. A high percentage of experts also agreed that 'a lack of incentive to disseminate findings' (64 %) and 'a lack of an accepted standard' (69 %) were also valid explanations for gaps in the evidence base. For example, one panellist stated that:

> In most conservation organisations there is little time or incentive for staff to write up findings for journals; our data and analyses usually go no further than donor reports and project institutional databases, rather than reaching the scientific literature.

[INSERT TABLE 3]

Table 4 summarises the aggregate results from three final attitude questions posed to the panel in R2. It was found that there was a strong consensus regarding how important it was 'to develop an accepted standard for the design and implementation of conservation evaluations,' with 85 % of the expert panel concurring that this development would be very important or at least quite important (Table 4). In contrast, there was found to be less agreement (only 61%) amongst the panel when asked to what extent they agreed that formal impact evaluation methods are unsuitable for evaluating *all* types of conservation policies.

[INSERT TABLE 4]

Finally, panellists were directly asked for their opinion on whether sufficient effort was being made in their organisation to improve programme evaluation standards. Encouragingly, 50 %, of the panel member answered 'Yes,' (Table 4).

5. Discussion and policy recommendations

The first point of inquiry of this study was to more systematically investigate to what extent rigorous evaluation methods are being used within conservations organisations. In line with existing arguments in the literature, the results from the Delphi study confirmed that the majority of evaluations in the sample of conservation organisations studied still use a simple before-after or with-without design (Figure 1). However, while the results show that the use of more rigorous evaluation methods is still insufficient in the conservation policy field, it does appear to be less limited within our sample than previously suggested. In fact, the common view held by the panel is that there has been an increase in the application of these methods in recent years and several panellists were aware of some organisations that are already regularly taking a more rigorous approach. For instance, many of the panel drew attention to a significant programme of work administered by conservation organisations (such as the RSPB, Natural England, CIFOR and WWF) that use BACI experiments and quasi-experimental methods to evaluate the environmental and social impacts of conservation interventions. This body of work on account of being either unpublished or in grey literature has often not been picked up by past critical reviews. Although not all of these studies can strictly be said to be perfect examples of impact evaluations, they do demonstrate that these organisations are making an attempt to construct reasonably credible counterfactuals. That said, given that the same few examples were provided by the panel, it is apparent that these organisations are still the exception and not the rule.

While the results of this study corroborate previous claims that the use of experimental and quasi-experimental methods is still limited, they do not, however, support assertions that this is largely because conservationists are ignorant and unappreciative of rigorous methods. In contrast, the results support that the panel were not only highly aware of these methods and the need for more credible evidence but also recognised their importance when it came to drawing reliable inferences about the causal effects of conservation interventions (Table 2). What is more, far from being ignorant of experimental and quasi-experimental evaluation methods, the consensus reached amongst the panel was that these methods are, at least in theory, considered to be the benchmark, or 'best practice' for conservation evaluation (Figure 1).

In line with Roe, Greig-gran & Mohammed's (2013) arguments, the results reveal that there is a considerable gap between what methods and design considerations are considered to be 'best practice,' and what methods are actually feasible to implement in reality. While there is far from a simple explanation for *why* this implementation gap persists, as discussed above it is clearly not because impact evaluation in not valued in the conservation policy field (Table 3). Instead, the results from this study show that there are in fact a number of pervasive barriers to implementing experiments and quasi-experiments on the ground. In particular, many of the experts felt that the use of rigorous evaluation methods would remain limited without more staff and programme consistency and substantially higher levels of funding to implement evaluations over longer time periods.

The 'crisis management nature' of the conservation policy field (Pullin et al., 2013) was

also reflected in many of the comments left by the panel who argued that the forward planning required for rigorous evaluations is often just too impractical in the face of short-term funding and thus short-lived opportunities for action which are urgently needed. For the same reasons, our study also showed that there is a lack of incentive for conservation organisations to disseminate their findings.

Whilst a lack of data sharing is not a barrier to evaluation *per se*, the vast majority of the panel agreed that it was likely to be one of the reasons for the gaps in the evidence base and is therefore an area in need of improvement (Table 3). What is more, the finding that much of the data in conservation does not reach the scientific literature supports the theory that the reliance on desk reviews of the scientific literature in order to assess approaches to evaluation in conservation is unlikely to be a fair or accurate assessment of what is happening in reality and, thus, the results may be a gross underestimate of the progress in evaluation that is actually taking place. For instance, a review of the Rainforest Alliance's unpublished impact studies, which was conducted as an extension to this study, found that there was mounting evidence of counterfactual thinking within the conservation organisation as many of their more recent evaluations had at least attempted to construct reasonably credible counterfactuals using matching methods (e.g., Paschall & Seville, 2012 and Hughell & Newsom, 2013). Unfortunately, this progress appears to have been overlooked by some of the critical literature which has largely focused on material published in more academic oriented sources or only readily available reports.

Further, many experts agreed that experimental and quasi-experimental methods should be prescribed in a more targeted manner. These sentiments are synonymous with arguments recently made by Pullin and his colleagues (2013), which stressed that while evaluation was important, it was also time consuming and costly and therefore needed to be justified. Furthermore, Mascia and others (2014), as well as Roe, Greig-gran & Mohammed (2013), have suggested experimental and quasi-experimental methods are best employed selectively where additional rigour is required to inform major programme decisions.

Finally, one of the key contributions of this study relates to the methods used. Whilst there has been much discussion on standards of evaluation methods used by the conservation sector, the reasons provided for the current trends and attitudes towards these methods have largely been based on personal experience and anecdotal evidence. In contrast, this study is the first to employ the Delphi technique to provide a more systematic assessment of conservation experts' actual knowledge, appreciation, as well as their experience with such methods. Overall, despite its limitations, the Delphi method proved to work well. Throughout the study, experts were observed to react to information from other experts and a considerable amount of convergence was observed to produce a clear consensus on a number of issues. Further the sampling frame adopted (including over 1600 conservation experts) as well as the actual sample size of respondents (n=45) are considerably above the minimum requirements for achieving robust findings (Hasson et al., 2000). Lastly, respondent bias was minimised as experts were rigorously selected to ensure that a wide spectrum of organisations and expertise were represented (Table 1). That said, it is important not to over generalise the findings

of this study; the results are only representative of the conservation organisations in the sampling frame and cannot be considered to be necessarily representative of the conservation sector at large or of a specific geographical context.

The insights gained from of our Delphi study allow us to draw several policy recommendations with respect to the use of impact evaluation methods in conservation. Firstly, our analysis suggests that the focus of research should move away from codifying best practice evaluation methods and instead focus on developing and improving minimum standards. As such, more emphasis should be placed on getting the basics of evaluation right. Indeed, there was a strong consensus amongst the panel that it was particularly important to first develop an accepted standard for the design and implementation of conservation evaluations (Table 3) as the current lack of an accepted standard was considered to be another factor contributing to gaps in the evidence base (Table 4). The standards would include the requirement that baseline survey data (before the intervention) on the environmental and livelihood impacts of conservation policies are more routinely collected from both potentially treated and comparable untreated (or control) villages. By agreeing to such basic minimum standards of policy evaluation, policy organisations position themselves more favourably in order to undertake evaluation in the future (when many of the impacts of their policies will be more readily observable).

Secondly, it should be acknowledged that not all conservation policies can or should be subjected to large-scale rigorous policy impact evaluation. Policy agencies on their own will unlikely have the capacity and know-how to independently design and implement such studies. Aiming to undertake a plethora or of ill-designed evaluation studies will not provide valuable information and will constitute a waste of time and money. Instead what *is* needed is the emergence of a selected critical mass of carefully designed and executed evaluation studies (including RCTs and framed field experiments) that will produce unbiased estimates of impacts of conservation policies across different geographical and institutional contexts. This will enable researchers to undertake meta-analyses of this type of unbiased evidence that will produce more generalisable findings.

Thirdly, undertaking such detailed and rigorous impact evaluation studies requires considerable time commitment (often over several years) as well as funds and resources not available to conservation organisations. Hence for this purpose it is imperative that NGOs collaborate with academics and gain access to additional resources to complete such studies. Research grant agencies should more proactively support and facilitate such collaborative research projects. A paradigm example of such a collaboration is that between the Royal Society for the Protection of Birds (RSPB) and academics from Wageningen and Cambridge universities, whom, between 2010-2015, undertook a series of comprehensive and rigorous policy impact evaluation studies (including randomised control trials) on the environmental and social impacts of conservation policies that aim to preserve the Gola Forest Nature Reserve in Sierra Leone (see project link here: http://www.conservation.cam.ac.uk/collaboration/framework-assessing-livelihood-impacts-forest-conservation-programmes). It is imperative that these types of projects are emulated and funded by research grant agencies.

Fourth, the funders of conservation projects themselves also need to change their priorities and adopt a culture in which conservation evaluation is given as much importance as conservation action. This change in attitude is essential for providing the incentive to conservation practitioners to undertake or, at a minimum, engage with impact evaluation studies (Stravinsky et al., 2000; Kapos et al., 2008). For example, national and international policy organisations as well as private market stakeholders that are involved in the funding of Reducing Emissions from Deforestation and Forest Degradation (REDD) projects or Payments for Ecosystem Services (PES) programmes should embrace the importance and necessity of impact evaluation by providing adequate time and resources to undertake such studies and disseminate the results obtained.

6. Conclusion

The aim of this study was to more formally and systematically assess the importance and use of rigorous evaluation methods in the conservation policy field by conducting a Delphi survey of conservation experts with real experience in the conservation sector. Using a Delphi technique proved to be an effective way of synthesising expert knowledge to produce a coherent and comprehensive picture of the rationale behind conservation evaluation decisions in practice and, thus, has provided important insights into each of the study's five research questions.

In general, the results confirm that the use of experimental and quasi-experimental evaluation methods in conservation is still very limited but this, crucially, is not because conservationists are ignorant of these methods or do not recognise them as being superior to non-experimental methods. In fact, considerable effort is being made to develop and improve evidence standards but these efforts have largely been thwarted by large financial and time related constraints that mean even elementary evaluations are hard to achieve. Impact evaluation is clearly not a panacea and will not always be what is needed. Certainly, incessant calls for increasingly rigorous evaluations are likely both quixotic and unproductive. Instead, this study recommends that there should be less focus in the literature on codifying best practice and more focus on finding ways to effectively raise minimum standards on a tight budget. Further, state-of-the-art impact evaluations should be aimed for a small selected number of case studies.

As the Delphi study has proved to be an effective communication device in this area, a way forward from this study could be to widen the scope of this Delphi study to incorporate the views of the academic community by adding another panel comprised of conservationists working purely in research and academia. This way a discussion between the practitioners and researchers could be facilitated in an attempt to identify common views, share knowledge and seek ever more efficient means to a common end. Whatever the method employed, explaining, and thus addressing, the gaps in the evidence base will require academics to put their prejudices aside, open up paths of communication with the conservation sector and, crucially, undertake more research that draws on the expert knowledge and experience of those on the front line of conservation practice.

Supplementary Materials

Supplementary Materials related to this article are attached as Appendixes SM1 to SM4 (and will be made available on the JEMA's website). These include a Technical Annex detailing the methods (Appendix SM1), the briefing document sent to the panellists (Appendix SM2), and copies of the survey instruments used for round one and round two of the survey (Appendix SM3 and Appendix SM4, respectively). The authors are solely responsible for the content of these materials. Queries (other than absence of material) should be directed to the corresponding author.

Acknowledgements

We would like to extend particular thanks to the 45 panellists who participated in our Delphi study for their engagement and expertise, as well as three anonymous reviewers for their insightful comments on an earlier draft of this article which greatly improved the manuscript. We would also like to thank Dr. Valerie Kapos for testing the survey instruments and providing useful feedback.

References

Adhikari, B., Agrawal, A., 2013. Understanding the Social and Ecological Outcomes of PES Projects: A Review and an Analysis. Conservat. Soc. 11, 359-374. http://dx.doi.org/10.4103/0972-4923.125748

Alcorn, J.B., 2014. Lessons learned from Community Forestry in Latin America and their relevance for REDD+. USAID-supported Forest Carbon, Markets and Communities (FCMC) Program, EUA. http://www. fcmcglobal. org/documents/CF_Latin_America. pdf. Accessed 31 March 2016.

Angus, A., Hodge, I., McNally, S., Sutton, M., 2003. The setting of standards for agricultural nitrogen emissions: a case study of the Delphi technique. J. Environ. Manag. 69, 323-337. http://dx.doi.org/10.1016/j.jenvman.2003.09.006

Baylis, K., Honey-Rosés, J., Börner, J., Corbera, E., Ezzine-de-Blas, D., Ferraro, P.J., Lapeyre, R., Persson, U.M., Pfaff, A., Wunder, S., 2015. Mainstreaming Impact Evaluation in Nature Conservation. Conserv. Lett. 9, 58-64. http://dx.doi.org/ 10.1111/conl.12180

Berry, M., Cashore, B., Clay, J., Fernandez, M., Lebel, L., Lyon, T., Mallet, P., 2012. Toward sustainability: The roles and limitations of certification. Resolve, Inc. Washington, DC:

Blackman, A., Rivera, J.E., 2010. The evidence base for environmental and socioeconomic impacts of 'sustainable'certification. Resources for the Future, Washington, DC. https://www.researchgate.net/profile/Jorge_Rivera10/publication/46456069_The_Evidence_Base_for_Environmental_and_Socioeconomic_Impacts_of_Sustainable_Certificati on/links/0c9605299683547c30000000.pdf Accessed 11 April 2016.

Blackman, A. (2012). Expost evaluation of forest conservation policies using remote sensing data: An introduction and practical guide. EfD, Discussion Paper Series 12-05. http://www.rff.org/RFF/Documents/EfD-DP-12-05.pdf Accessed May 19 2015.

Bottrill, M.C., Hockings, M., Possingham, H.P., 2011. In pursuit of knowledge: addressing barriers to effective conservation evaluation. Ecol. Soc. 16, 14 http://www.ecologyandsociety.org/vol16/iss2/art14/

Cowling, R.M., 2014. Let's Get Serious About Human Behavior and Conservation. Conserv. Lett. 7, 147-148. http://dx.doi.org/10.1111/conl.12106

Duflo, E., Glennerster, R., Kremer, M., 2008. Using randomization in development economics research: A toolkit, in: Schultz, T. and Strauss, J. (Eds.), Handbook of Development Economics., 4. New York, pp. 3895-3962.

Ferraro, P.J., Pattanayak, S.K., 2006. Money for nothing? A call for empirical evaluation of biodiversity conservation investments. PLoS Biol. 4, e105. http://dx.doi.org/10.1371/journal.pbio.0040105

Ferraro, P.J., 2009. Counterfactual thinking and impact evaluation in environmental policy, in: Birnbaum, M. & Mickwitz, P., (Eds), Environmental Program and Policy Evaluation: Addressing Methodological Challenges. New Directions for Evaluation 122, pp. 75-84.

Fisher, B., Balmford, A., Ferraro, P.J., Glew, L., Mascia, M., Naidoo, R., Ricketts, T.H., 2014. Moving Rio Forward and Avoiding 10 More Years with Little Evidence for Effective Conservation Policy. Conserv. Biol. 28, 880-882. http://dx.doi.org/10.1111/cobi.12221

Frondel, M., Schmidt, C.M., 2005. Evaluating environmental programs: The perspective of modern evaluation research. Ecol. Econ. 55, 515-526. http://dx.doi.org/10.1016/j.ecolecon.2004.12.013

Geneletti, D., 2008. Incorporating biodiversity assets in spatial planning: Methodological proposal and development of a planning support system. Landscape Urban Plan. 84, 252-265. http://dx.doi.org/10.1016/j.landurbplan.2007.08.005

Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B., Vermeersch, C.M., 2011. Impact Evaluation in Practice. World Bank Publications.

Greenstone, M and Gayer, T., 2009 Quasi-Experimental and Experimental Approaches to Environmental Economics. J. Environ. Econ. Manag. 57, 21-44. http://dx.doi.org/10.1016/j.jeem.2008.02.004

Hasson, F., Keeney, S., McKenna, H., 2000. Research guidelines for the Delphi survey technique. J. of Adv. Nurs. 32, 1008-1015. http://dx.doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x

Hess, G.R., King, T.J., 2002. Planning open spaces for wildlife: I. Selecting focal species using a Delphi survey approach. Landscape Urban Plan. 58, 25-40. http://dx.doi.org/10.1016/S0169-2046(01)00230-4

Hughell, D. & Newsom, D., 2013. Evaluating the results of our work: impacts of Rainforest Alliance certification on coffee farms in Colombia. Cenicafe, Colombia. http://www.rainforest-alliance.org/sites/default/files/publication/pdfcenicafe_singles_0.pdf. Accessed 13 July 2014.

Kapos, V., Balmford, A., Aveling, R., Bubb, P., Carey, P., Entwistle, A., Hopkins, J., Mulliken, T., Safford, R., Stattersfield, A., Walpole, M., Manica, A., 2008. Calibrating

conservation: new tools for measuring success. Conserv. Lett. 1, 155-164. http://dx.doi.org/10.1111/j.1755-263X.2008.00025.x

Khandker, S.R., Koolwal, G.B., Samad, H.A., 2010. Handbook on Impact Evaluation: Quantitative Methods and Practices. World Bank Publications.

Kleiman, D.G., Reading, R.P., Miller, B.J., Clark, T.W., Scott, J.M., Robinson, J., Wallace, R.L., Cabin, R.J., Felleman, F., 2000. Improving the Evaluation of Conservation Programs. Conserv. Biol. 14, 356-365. http://dx.doi.org/ 10.1046/j.1523-1739.2000.98553.x

Linstone, H.A., Turoff, M., 1975. The Delphi Method: Techniques and Applications. Addison-Wesley Reading, MA.

MEA (Millenium Ecosystem Assessment), 2005. Ecosystems and Human Well-being. Island Press Washington, DC.

MacMillan, D.C., Marshall, K., 2006. The Delphi process – an expert-based approach to ecological modelling in data-poor environments. Animal. Conserv. 9, 11-19. http://dx.doi.org/10.1111/j.1469-1795.2005.00001.x

Mascia, M.B., Pailler, S., Thieme, M.L., Rowe, A., Bottrill, M.C., Danielsen, F., Geldmann, J., Naidoo, R., Pullin, A.S., Burgess, N.D., 2014. Commonalities and complementarities among approaches to conservation monitoring and evaluation. Biol. Conserv. 169, 258-267. http://dx.doi.org/10.1016/j.biocon.2013.11.017

McKinnon, M.C., Mascia, M.B., Yang, W., Turner, W.R., Bonham, C., 2015. Impact evaluation to communicate and improve conservation non-governmental organization performance: the case of Conservation International. Philos. T. Roy. Soc. B. 370. http://dx.doi.org/10.1098/rstb.2014.0282

Miteva, D.A., Pattanayak, S.K., Ferraro, P.J., 2012. Evaluation of biodiversity policy instruments: what works and what doesn't? Oxf. Rev. Econ. Pol. 28, 69-92. http://dx.doi.org/10.1093/oxrep/grs009

Okoli, C., Pawlowski, S.D., 2004. The Delphi method as a research tool: an example, design considerations and applications. Inform. Manag. 42, 15-29. http://dx.doi.org/10.1016/j.im.2003.11.002

Oliver, I., 2002. An expert panel-based approach to the assessment of vegetation condition within the context of biodiversity conservation: Stage 1: the identification of condition indicators. Ecol. Indic. 2, 223-237. http://dx.doi.org/10.1016/S1470-160X(02)00025-0

Orsi, F., Geneletti, D., Newton, A.C., 2011. Towards a common set of criteria and indicators to identify forest restoration priorities: An expert panel-based approach. Ecol. Indic. 11, 337-347. http://dx.doi.org/10.1016/j.ecolind.2010.06.001

Paschall, M. & Seville, D. 2012. Certified cocoa: scaling up farmer participation in West Africa. New Business Models for Sustainable Trading Relationships: IIED, CIAT, Rainforest Alliance, CRS & Sustainable Food Lab. http://pubs.iied.org/pdfs/16034IIED.pdf . Accessed 7 July 2014.

Pattanayak, S.K., Wunder, S., Ferraro, P.J., 2010. Show me the money: do payments supply environmental services in developing countries? Rev. Environ. Econ. Policy. 4, 254-274. http://dx.doi.org/10.1093/reep/req006

Pullin, A.S., Knight, T.M., 2001. Effectiveness in conservation practice: pointers from medicine and public health. Conserv. Biol. 15, 50-54. http://dx.doi.org/10.1111/j.1523- 1739.2001.99499.x

Pullin, A.S., Knight, T.M., Stone, D.A., Charman, K., 2004. Do conservation managers use scientific evidence to support their decision-making? Biol. Conserv. 119, 245-252. http://dx.doi.org/10.1016/j.biocon.2003.11.007

Pullin, A.S., Sutherland, W., Gardner, T., Kapos, V., Fa, J.E., 2013. Conservation priorities: identifying need, taking action and evaluating success, in: Macdonald, D.W., Willis, K. (Eds), Key Topics in Conservation Biology 2, John Wiley Publishing, pp. 3- 22.

Roe, D., Grieg-Gran, M., Mohammed, E.Y., 2013. Assessing the social impacts of conservation policies: rigour versus practicality. IIED Briefing Papers. International Institute for Environment and Development, London. http://pubs.iied.org/pdfs/17172IIED.pdf. Accessed 11 April 2016.

Rossi, P.H., Lipsey, M.W., Freeman, H.E., 2003. Evaluation: A systematic approach. Sage publications.

Salafsky, N., Margoluis, R., Redford, K.H., Robinson, J.G., 2002. Improving the practice of conservation: a conceptual framework and research agenda for conservation science. Conserv. Biol. 16, 1469-1479. http://dx.doi.org/10.1046/j.1523- 1739.2002.01232.x

Salafsky, N., Margoluis, R., 2003. What conservation can learn from other fields about monitoring and evaluation. BioSci. 53, 120-122. http://dx.doi.org/10.1641/0006- 3568(2003)053[0120:wcclfo]2.0.co;2

Samii, C., Lisiecki, M., Kulkarni, P., Paler, L., Chavis, L., 2014. Effects of payments for environmental services (PES) on deforestation and poverty in low and middle in- come countries: a systematic review. Cambell. Syst. Rev. 11. http//dx.doi.org/10.4073/csr.2014.11

Stem, C., Margoluis, R., Salafsky, N., Brown, M., 2005. Monitoring and evaluation in conservation: a review of trends and approaches. Conserv. Biol. 19, 295-309.

http://dx.doi.org/ 10.1111/j.1523-1739.2005.00594.x

Stevens, A., Abrams, K., Brazier, J., Fitzpatrick, R., Lilford, R., 2001. The Advanced Handbook of Methods in Evidence Based Healthcare. Sage, London.

Stravinsky, I., 2000. Writing the wrongs: developing a safe-fail culture in conservation. Conserv. Biol. 14, 1567-1568.

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends. Ecol. Evolut. 19, 305-308. http://dx.doi.org/10.1016/j.tree.2004.03.018

White, H., 2006. Impact evaluation: the experience of the Independent Evaluation Group of the World Bank. University Library of Munich, Germany. https://mpra.ub.uni- muenchen.de/1111/1/MPRA_paper_1111.pdf . Accessed 11 April 2016.

Zheng, H., Robinson, B.E., Liang, Y.-C., Polasky, S., Ma, D.-C., Wang, F.-C., Ruckelshaus, M., Ouyang, Z.-Y., Daily, G.C., 2013. Benefits, costs, and livelihood implications of a regional payment for ecosystem service program. Proc. Natl. Acad. Sci. USA. 110, 16681-16686.

Tables and Figures

Table 1. Composition of the Delphi Panel in R1 and R2

	R1	**R2**
No. Experts Invited	200	45
No. Responses	45	36
Response Rate (%)	24[a]	80
Responses Female (%)	45	47
Responses Male (%)	55	53
No. Organisations Represented	24[b]	21[c]

[a]adjusted to account for 5 % of emails bouncing back, i.e.,190 sent and received in total.

[b]including: WWF-US, UNEP-WCMC, WCS, UNESCO, RSPB, FFI, CCI, The Natural Capital Initiative, John Muir Trust, Bioversity International, Rainforest Alliance, Endangered Wildlife Trust, Natural England, DEFRA, FERA, FAO, OECD, FC, IUCN, IIED, ICL, DICE, Institute for Forest and Environmental Policy, Cambridge Forum for Sustainability and the Environment.

[c]less: OECD, John Muir Trust and Bioversity International.

Table 2. Aggregate scores and responses for seven attitude questions posed to the Delphi panel in R1.

Research Question Being Addressed	Question (presented verbatim)	Median	Mean	*S.D*	*n*	Percentage of Respondents
RQ1	Efforts to evaluate the effectiveness of conservation interventions are essential to building the evidence base of what works and when. **1 = Strongly Agree, 2 = Agree, 3 = Neither/Nor, 4 = Disagree, 5 = Strongly Disagree**	1	1.24	0.48	45	98 % Strongly Agree or Agree
RQ1	How important is the use of experimental and quasi-experimental programme evaluation methods for drawing reliable inferences about the causal effects of conservation interventions? **1 = Very, 2 = Quite, 3 = Neither/Nor, 4 = Not Important, 5 = Completely Inappropriate**	2	1.76	0.61	45	95 % rated Very Important or Quite Important
RQ3	When it comes to evaluating the success of its interventions, the field of ecosystem and biodiversity conservation lags behind most other policy fields. **1 = Definitely yes, 2 = Probably yes, 3 = Probably not, 4 = Definitely not**	2	2.04	0.77	45	77 % rated Definitely or Probably yes
RQ3	How often are attempts made to measure the outcomes and impacts of conservation interventions using programme evaluation methods? **1 =Always, 2 = Usually, 3 = Occasionally, 4 = Never**	3	2.82	0.45	44	77 % said Occasionally
RQ3	On average what proportion of these evaluation studies would you say use experimental or quasi-experimental designs? **1 = All, 2 = Most, 3 = Some, 4 = None**	3	3.05	0.45	40	80 % said Some
RQ3	How would you best describe the general trend in the use of experimental and quasi-experimental evaluation methods in the conservation policy field over the last few years? **1 = Dramatic Increase, 2 = Moderate Increase, 3 = Slight Increase, 4 = No Change, 5 = Slight Decline, 6 = Moderate Decline, 7 = Dramatic Decline**	3	2.69	0.83	39	77 % said Moderate Increase or Slight Increase
RQ5	In general, conservation organisations are working hard to improve their programme evaluation standards in an attempt to strengthen the credibility of the evidence base. **1 = Strongly Agree, 2 = Agree, 3 = Neither/Nor, 4 = Disagree, 5 = Strongly Disagree**	2	2.40	0.75	45	58 % Strongly Agree or Agree

Table 3. Potential explanations for gaps in the evidence base: the results presented here inform RQ4 (from R2 survey)

To what extent do you agree or disagree with each of the following explanations for any gaps in the evidence base regarding the impacts of conservation interventions? **1 = Strongly Agree, 2 = Agree, 3 = Neither/Nor, 4 = Disagree, 5 = Strongly Disagree**					
Statement (presented verbatim)	**Median**	**Mean**	**S.D**	**n**	**Percentage of Respondents Agreeing or Strongly Agreeing**
Gaps in the evidence base have less to do with the nature of the field and more to do with a lack of incentives and/or funding/resources.	2	2.14	0.83	36	75 %
Gaps in the evidence base can mainly be attributed to a lack of funding and/or resources and not because impact evaluation is not valued in the conservation policy field.	2	2.17	0.56	36	75 %
Gaps in the evidence base can partly be explained by a lack of incentive to disseminate findings: writing-up results for journals in not a priority	2	2.47	0.94	36	64 %
Gaps in the evidence base can partly be explained by the lack of an accepted standard for the design and implementation of impact evaluations in the conservation policy field	2.5	2.67	1.01	36	69 %

Table 4. Attitudes towards impact evaluation (from R2 survey)

Research Questions Being Addressed	Question (presented verbatim)	Median	Mean	*S.D*	*n*	Percentage of Respondents
RQ2	Experimental (randomised evaluations) and quasi-experimental (statistical matching) methods are not suitable for evaluating all conservation interventions and should only be used in certain circumstances. **1 = Strongly Agree, 2 = Agree, 3 = Neither/Nor, 4 =**	2	2.47	1.00	36	61 % Strongly Agree or
RQ4	In your opinion, how important or unimportant is it to develop an accepted standard for the design and implementation of conservation evaluations? **1 = Very, 2 = Quite, 3 = Neither/Nor, 4 = Not Important, 5 =**	2	2.06	0.79	36	83 % rated Very Important or Quite Important
RQ5	With regard to your own organisation, do you think sufficient effort is being made to develop or improve programme evaluation standards in an attempt to strengthen the credibility of the evidence base? **1 = Yes, 2 = No, 3 = Don't Know**	1.5	1.78	0.87	36	50 % said Yes

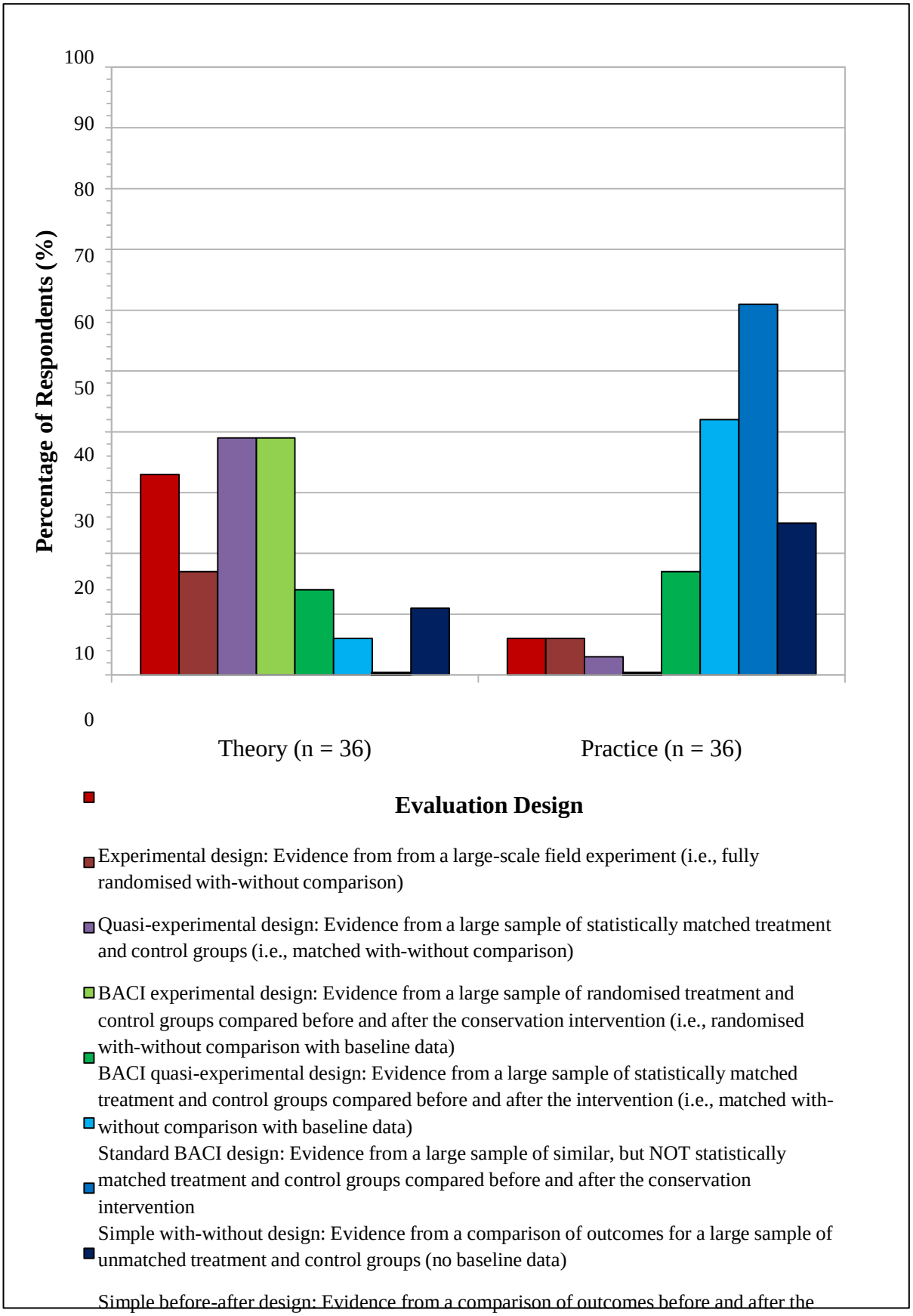

Figure 1. Comparison of the evaluation design/s that were chosen by the panel in R2 when asked what they considered to be the ideal evaluation design in theory and the most commonly used design in practice. Panellists could choose one or two answers from the eight options available for

each question. The results presented here inform RQ2 and RQ3

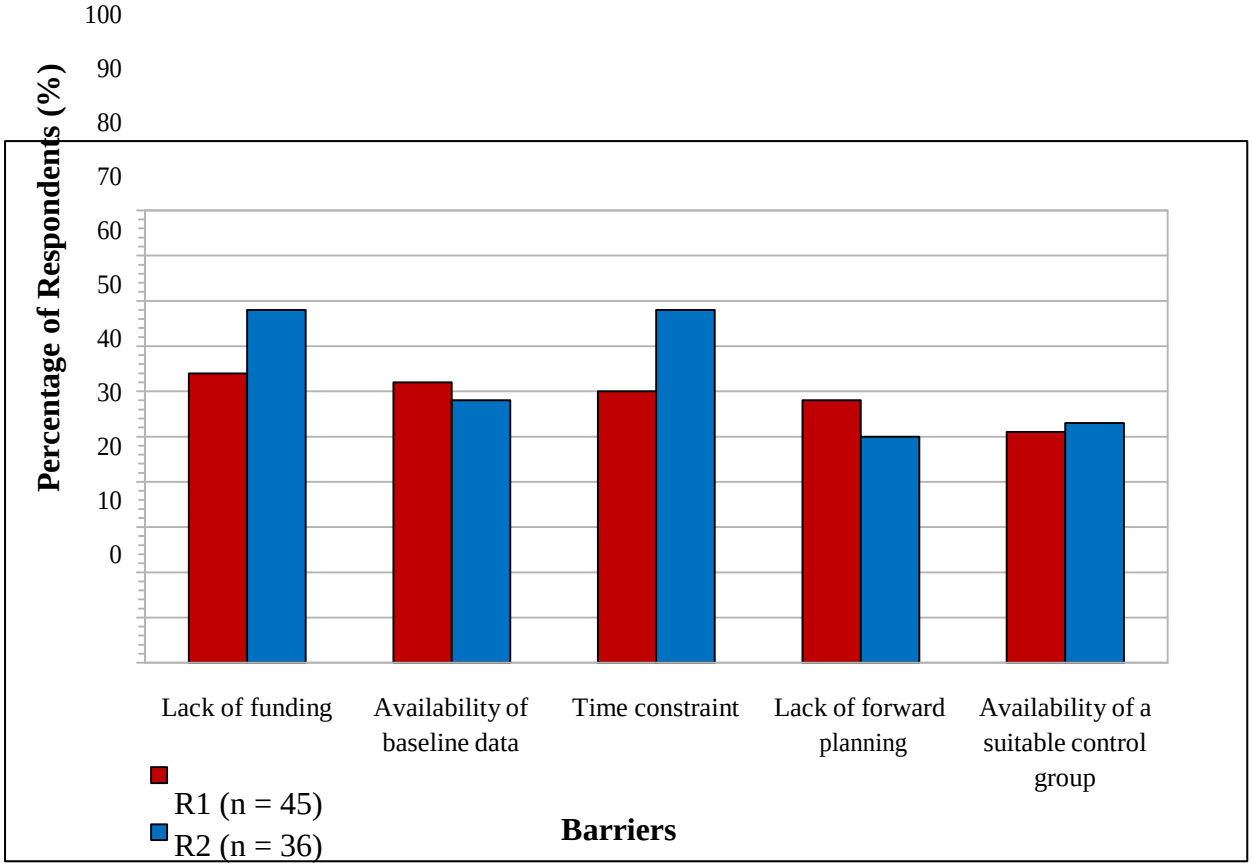

Figure 2. The five most suggested barriers to implementing experimental and quasi-experimental evaluation methods in the conservation policy field as rated by the Panellists in R1 and then again in R2. The results presented here inform RQ4.